DOCUMENT RESUME

ED 334 210 TM 016 631

AUTHOR Luk, HingKwan
TITLE An Empirical Comparison of an Expert Systems Approach and an IRT Approach to Computer-Based Adaptive Mastery Testing.
PUB DATE Apr 91
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; *Adaptive Testing; College Entrance Examinations; Comparative Analysis; *Computer Assisted Testing; *Estimation (Mathematics); *Expert Systems; Higher Education; Item Response Theory; *Language Tests; *Mastery Tests; Mathematical Models; Simulation; *Spanish; *Undergraduate Students
IDENTIFIERS College Board Achievement Tests; *EXSPRT; Placement Tests; Three Parameter Model

ABSTRACT

This study examined whether an expert system approach involving intelligent selection of items (EXSPRT-I) is as efficient as item response theory (IRT) based three-parameter adaptive mastery testing (AMT) when there are enough subjects to estimate the three IRT item parameters for all items in the test and when subjects in the item parameter estimation (IPE) sample are different from those in the validation sample. EXSPRT is based on expert systems reasoning, Bayesian reasoning, and A. Wald's sequential probability ratio test. The 100-item multiple-choice College Entrance Examination Board Spanish Achievement Test (CEEBSAT) was administered to 1,672 undergraduate students at Indiana University (Bloomington): 1,000 examinees were assigned to an IPE group, and 672 examinees were assigned to a validation group. The CEEBSAT is a paper-and-pencil test consisting of a 40-item listening subtest and a 60-item reading subtest. IPE and test reenactments were conducted separately for the two subtests. The two subject groups were essentially the same with respect to the two subtests. Although EXSPRT-I decision accuracies were not as high as expected, they were reasonably high for many practical purposes. EXSPRT-I required the least number of test items to reach a mastery or non-mastery decision for non-masters, followed by AMT, and then EXSPRT with random selection of items. EXSPRT-I is more accurate and more efficient in making mastery decisions than AMT; however, it is less accurate but more efficient than AMT for non-mastery decisions. EXSPRT is less complicated conceptually and mathematically and is more practical since it does not require a large parameter estimation sample. Three data tables and a 13-item list of references are included. Study formulas and algorithms are presented. (RLC)

ED334210

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

HingKwan Luk

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

An Empirical Comparison of an Expert Systems Approach and an IRT Approach to Computer-Based Adaptive Mastery Testing

HingKwan Luk
Bureau of Evaluative Studies and Testing
Indiana University
Bloomington

Paper presented at the Annual Meeting of the
American Educational Research Association

April, 1991

TM016631

BEST COPY AVAILABLE

Abstract

Researchers have investigated different methodologies for computerized adaptive mastery testing. One methodology that has been much examined is the item response theory (IRT). However, IRT requires a large number of examinees for estimation of item parameters (a minimum of 200 to 1000, depending on the number of item parameters to be estimated and the number of items in the pool.) Frick, Plew, and Luk (1989) have investigated an expert systems approach, EXSPRT, to adaptive mastery testing. They demonstrated that EXSPRT was more efficient than, or as efficient as, IRT in three empirical studies. However, there were not enough subjects in their studies for estimation of all three IRT parameters. The present study was conducted to examine if EXSPRT is comparable to IRT when all three IRT parameters can be empirically estimated.

IRT-based Adaptive Mastery Testing

Item response theory (IRT) has been used in the development of computerized adaptive testing. One notable example is Kingsbury and Weiss' (1983) adaptive mastery testing (AMT), which incorporates both IRT and Bayesian reasoning to make a mastery test adaptive to the examinee's achievement level.

The AMT assumes a one-, two-, or three-parameter logistic ogive (an item characteristic curve) that describes the functional relationship between the probability of observing a correct response to a test item on the mastery test and a latent achievement (θ) metric. The three parameters are item discriminatory power (a_i), item difficulty (b_i), and lower asymptote (c_i—guessing parameter).

Using Birnbaum's (1968) formula (see Appendix 1), the three item parameters can be used to determine the amount of information that each test item provides regarding the examinee's θ estimate at a particular point in a test. Brown and Weiss (1977) incorporated this concept of item information in a "maximum information search and selection" (MISS) procedure. According to MISS, the item that provides the most information about the examinee's current θ estimate is selected and administered next. The examinee's response to the selected item is then scored, and a new θ estimate is made using Owen's (1975) Bayesian scoring algorithm (see Appendix 1). Owen's algorithm also gives an estimate of the variance of the examinee's

estimated θ. A confidence interval for the examinee's θ is therefore defined.

The confidence interval for the examinee's θ is then compared with a pre-determined θ cutoff (θ_c) to decide if testing is to continue. θ_c is the θ value corresponding to a pre-determined mastery cutoff in terms of proportion of items answered correctly. The conversion of a mastery cutoff on a proportion-correct scale to θ_c is based on a test characteristic curve that can be considered as an average of all item characteristic curves in a test (Weiss and Kingsbury, 1984).

If the lower bound of the confidence interval lies above θ_c, mastery is concluded. If the upper bound of the confidence interval lies below θ_c, a nonmastery decision is made. If the confidence interval includes θ_c, testing continues. The most informative item remaining in the item pool is then selected and administered, a new θ estimate is derived, and a new confidence interval is defined. The AMT procedure repeats itself until mastery or nonmastery is concluded, or until the item pool is exhausted. If the item pool is exhausted when the confidence interval still contains θ_c, a mastery or nonmastery decision will be made depending on whether the estimated θ is above or below θ_c.

It should be pointed out here that at the very beginning of the test administration, if nothing is known about the examinee's true θ, the θ estimate can be set to zero and the variance estimate can be set to one because ability scores (and hence θ's)

are assumed to be normally distributed with mean of zero and standard deviation of one (Hambleton and Cook, 1983).

One limitation of AMT is that it requires a relatively large number of subjects for reliable estimation of item parameters. As a rule of thumb, a minimum of 1000 subjects is required to reliably estimate a 3-parameter logistic ogive; and a minimum of 200 examinees for the one-parameter model. Educators and school teachers will often find it difficult, if not impractical, to test hundreds or thousands of subjects in advance when developing an adaptive mastery test.

EXSPRT: An expert systems approach

Frick, Plew, and Luk (1989) proposed EXSPRT as an alternative approach to adaptive mastery testing that does not require a large parameter estimation sample. EXSPRT is an approach based on expert systems reasoning, Bayesian reasoning, and Wald's (1947) sequential probability ratio test (SPRT). Wald's SPRT has been adapted by researchers (e.g., Ferguson, 1969; Frick, 1990; Kingsbury and Weiss, 1983) within a mastery testing situation.

Frick et al. (1989) conceived a logistic ogive as a set of expert systems rules in the form of "If the estimated ability (or θ) of an examinee is x, and if item y is selected, then the probability of the examinee answering the item correctly is z." However, such rules are more pertinent to a continuum of alternatives since the θ metric in IRT is a continuum. When an adaptive mastery test is used to choose only between a few

discrete alternatives such as mastery and nonmastery, other expert systems rules can be created. For example, a rule quadruplet can be constructed for each item *i* in an item pool:

Rule i.1: If the examinee is a *master* and item *i* is selected, then the probability of a correct response is $P(C_i|M)$

Rule i.2: If the examinee is a *master* and item *i* is selected, then the probability of an incorrect response is $P(\neg C_i|M)$

Rule i.3: If the examinee is a *nonmaster* and item *i* is selected, then the probability of a correct response is $P(C_i|N)$

Rule i.4: If the examinee is a *nonmaster* and item *i* is selected, then the probability of an incorrect response is $P(\neg C_i|N)$.

Such rule quadruplets, which are comparable to logistic ogives in IRT-based AMT, will form an EXSPRT rule base for an adaptive mastery test involving dichotomous decisions (mastery vs. nonmastery).

Rule quadruplets can be *empirically* derived by giving the pool of test items to a representative sample (the parameter estimation sample) of examinees who are supposed to show a wide range of test scores. After choosing an appropriate mastery cut-off score, probabilities of correct and incorrect responses for the mastery group (in the parameter estimation sample) for each item *i* in the pool can be computed using the following formulae:

$$P(C_i|M) = (nr_{im} + 1) \;/\; (nr_{im} + nw_{im} + 2) \qquad [1.1]$$

$$P(\neg C_i|M) = 1 - P(C_i|M) \qquad [1.2]$$

where nr_{im} = number of examinees in the mastery group who answered item i correctly;

and nw_{im} = number of examinees in the mastery group who answered item i incorrectly.

Similarly, for the nonmastery group, $P(C_i|N)$ and $P(\neg C_i|N)$ are defined for each test item i:

$$P(C_i|N) = (nr_{in} + 1) \;/\; (nr_{in} + nw_{in} + 2) \qquad [1.3]$$

$$P(\neg C_i|N) = 1 - P(C_i|M) \qquad [1.4]$$

where nr_{in} = number of examinees in the nonmastery group who answered item i correctly;

and nw_{in} = number of examinees in the nonmastery group who answered item i incorrectly.

In EXSPRT testing, Bayes' Theorem is applied to update the posterior probabilities of an examinee being a master or a nonmaster each time an examinee has responded to a test item. A likelihood ratio of the examinee being a master to being a nonmaster is then computed. The likelihood ratio is given by formula [2] below:

$$LR = \frac{P_{om} \prod_{i=1}^{n} [P(C_i|M)^{s} \; P(\neg C_i|M)^{f}]}{P_{on} \prod_{i=1}^{n} [P(C_i|M)^{s} \; P(\neg C_i|M)^{f}]} \qquad [2]$$

where

P_{om} = prior probability of the examinee being a master,

P_{on} = prior probability of the examinee being a nonmaster,

n = number of items in the pool,

$s = 1$, $f = 0$ if response to item i is correct,

$s = 0$, $f = 1$ if response to item i is not correct,

$s = 0$, $f = 0$ if item i is not given.

In accordance with Wald's SPRT decision rules, a mastery decision is made if LR of is greater than or equal to $(1 - \beta)/\alpha$, where α is the probability of a false mastery decision, and β is the probability of a false nonmastery decision. If LR is less than or equal to $\beta/(1 - \alpha)$, a nonmastery decision is made. If LR lies between $(1 - \beta)/\alpha$ and $\beta/(1 - \alpha)$, testing continues.

EXSPRT Item Selection

Item selection in EXSPRT can be random or "intelligent". EXSPRT involving random selection of test items is denoted as EXSPRT-R, and EXSPRT involving intelligent selection of items is denoted as EXSPRT-I.

Intelligent item selection is based on two indices derived from the rule quadruplets of the test items. The two indices are item discrimination and item/examinee incompatibility.

Item discrimination of test item i, denoted as D_i, is defined as the difference between the estimated probabilities of correct answers to item i by masters and nonmasters. Item discrimination of test item i can be expressed mathematically:

$$D_i = P(C_i|M) - P(C_i|N) \qquad [3]$$

Item/examinee incompatibility is an index of how well an item will match the estimated achievement or ability level of an examinee. Item/examinee incompatibility of item i with respect to examinee j, denoted as I_{ij}, is defined as the absolute value of

the difference between examinee j's estimated ability level (in terms of the proportion of correct answers j has given so far in the test) and the probability of an incorrect answer to item i by either a master or a nonmaster. Mathematically, I_{ij} is:

$$I_{ij} = |(1-P(C_i)) - E(\Phi_j)| \quad [4]$$

where $$P(C_i) = (nr_i + 1) / (nr_i + nw_i + 2) \quad [5]$$

$$E(\Phi_j) = (nr_j + 1) / (nr_j + nw_j + 2) \quad [6]$$

nr_i = number of people (masters **AND** nonmasters) in the parameter estimation sample who have answered item i correctly

nw_i = number of people (masters **AND** nonmasters) in the parameter estimation sample who have answered item i incorrectly

nr_j = number of items j has answered correctly thus far

nw_j = number of items j has answered incorrectly thus far

The ratio of D_i to I_{ij} is the utility index of item i for examinee j. The utility index, U_{ij}, is defined as:

$$U_{ij} = D_i / (I_{ij} + \delta) \quad [7]$$

where δ = some arbitrary small constant (e.g. 0.0000001) to prevent division by zero.

An item with a high utility index is one that discriminates well between masters and nonmasters and that is well compatible with j's current estimated ability level. In EXSPRT-I, the item

to be selected next is the one remaining in the item pool that has the highest utility index.

The Present Investigation

Frick et al. (1989) reported results from three empirical studies which demonstrated that EXSPRT-I was more efficient than, or as efficient as, the IRT-based AMT model. The three tests used in their studies were classroom tests and not surprisingly, there were not enough subjects and data for reliable estimation of all three IRT parameters. Only item difficulties were estimated in their studies. Furthermore, in two of the three studies, examinees selected for item parameter estimation were subsequently included again for validation.

The present study was conducted to investigate if EXSPRT, particularly EXSPRT-I, is as efficient as the IRT-based AMT when there are enough subjects to estimate the three IRT item parameters for all items in the test, and when subjects in the item parameter estimation sample are different from those in the validation sample.

Method

The Spanish Language Placement Test

The test used in this study was the College Entrance Examination Board (CEEB) Spanish Achievement Test. The test was used by the Spanish Department of Indiana University (IU) as a placement test. The test was a paper-and-pencil test that consisted of two parts: a 40-item listening subtest and a 60-item reading subtest. The listening subtest required examinees to

listen to a pre-recorded tape. The test was administered in group and the time allocated to the listening subtest was 20 minutes, while that to the reading subtest was 40 minutes. All the 100 items on the test were multiple choice items involving four alternatives. Based on the test responses of 1672 examinees in the present study, Cronbach's α was found to be 0.81 for the listening subtest, and 0.87 for the reading subtest.

Examinees

One thousand six hundred and seventy two undergraduate students at IU took the Spanish Achievement Test. One thousand examinees were randomly selected from the 1672 examinees. These 1000 examinees were assigned to an item parameter estimation group. Estimation of the three IRT item parameters and derivation of EXSPRT rule quadruplets were based on the actual test responses of these 1000 examinees.

The remaining 672 examinees were assigned to a validation group. Actual test responses of these 672 examinees were used to re-enact the Spanish placement test under each of the following three adaptive methodologies: AMT, EXSPRT-I, and EXSPRT-R.

Item Parameter Estimation

Since the listening subtest and the reading subtest were administered separately and measured different language skills, item parameter estimation and test re-enactments were done separately for each of the two subtests.

IRT parameters. A computer program, LOGIST 5 (Wingersky, Barton, and Lord, 1982), was used to estimate a_i's, b_i's, and c_i's

of items in the listening subtest and in the reading subtest separately. The program ran on an IBM mainframe (IBM 3090 System). Estimation was based on the actual test responses of the 1000 examinees in the item parameter estimation group. Since the placement test was administered in group as a paper-and-pencil test, it was not possible to differentiate omitted items from those that had not been reached by an examinee. Consequently, all items to which an examinee had not responded were treated as omitted items for the purpose of IRT parameter estimation.

EXSPRT parameters. The mastery cutoff scores were 60% for the listening subtest and 50% for the reading subtest. The 1000 examinees in the parameter estimation group were divided into masters and nonmasters of the two subtests on the basis of these two cutoff scores. Their actual responses to the listening test items and to the reading test items were used to construct the rule bases for the two subtests using formulae [1.1], [1.2], [1.3], and [1.4].

Test Re-enactments

A computer program written in C was developed by the author to re-enact the listening test and the reading test, using actual responses of the 672 examinees in the validation group, for each of the three adaptive methodologies: AMT, EXSPRT-I, and EXSPRT-R. The C program was a C version of the test re-enactment program that was used in Frick et al.'s (1989) studies but that was written in DAL (Digital Authoring Language). The current program

had been cross-validated carefully with the original DAL version. Also, the original DAL version had been cross-validated with another PASCAL version written independently by the second author of Frick et al.'s (1989) paper. The listening subtest and the reading subtest were re-enacted separately for each examinee in the validation group.

The mastery cutoff score was set to 0.6 for the listening subtest, and to 0.5 for the reading subtest. Each of these two cutoff scores corresponded approximately to the point of inflection on the respective test characteristic curve.

AMT re-enactment. At the beginning of each test re-enactment, the initial prior θ variance was set to 1 and the initial prior θ level was set to θ_c, that is, the θ value corresponding to the mastery cutoff score on the test characteristic curve (c.f. Weiss and Kingsbury, 1984). MISS was used as the item selection procedure. Based on whether the selected item was actually answered correctly or incorrectly by the examinee, a new estimate of the examinee's true θ was determined using Owen's algorithm to determine if another item was to be selected, or if mastery or nonmastery could be concluded using a 0.95 confidence interval. If all items in the test had been "administered" and no mastery or nonmastery conclusion could be reached, the examinee would be classified as a master if the estimated θ at the end of the test was greater than or equal to θ_c; otherwise the examinee would be considered as a nonmaster.

At the end of each test re-enactment, the program would record the number of items necessary to reach a mastery or nonmastery decision by the AMT model. The program also recorded if the AMT decision agreed with the total subtest decision, that is, the decision based on the examinee's actual score on the subtest.

EXSPRT-I re-enactment. Items were selected on the basis of their utility indices (see formulae [3] to [7]). The actual response of the examinee to the selected item was used to update the likelihood ratio, *LR* (see formula [2]). P_{om} and P_{on} were both set to 0.5; that is, it was assumed that there was a 50-50 percent chance that the examinee was a master or a nonmaster at the beginning of the test. Wald's SPRT decision rules were then applied to determine if testing was to continue. The values of α and β were both set to 0.025. In other words, the overall decision error rate was 0.05, thus making EXSPRT decision comparable to AMT decision based on 0.95 confidence interval.

As in the AMT re-enactment, the number of items selected and "administered" was recorded. The agreement between the EXSPRT decision and the total subtest decision was also recorded. Forced decisions were also made, as in AMT, when all test items were exhausted with no EXSPRT decision made. Mastery would be concluded if the proportion of correct responses by the examinee was equal to or above the mastery cutoff; otherwise nonmastery would be concluded.

EXSPRT-R re-enactment. The same procedures of EXSPRT-I re-enactments were followed except that items were selected randomly instead of intelligently on the basis of utility indices.

Results and Discussion

Table 1 shows the means and standard deviations of listening and reading scores for both parameter estimation group and validation group. T-tests were conducted to test the differences between means of listening scores in the parameter estimation group and those in the validation group. None of the pairwise

Table 1

Means and Standard Deviations of Test Scores

Group	Listening Test N Ss	Mean	S.D.	Reading Test N Ss	Mean	S.D.
Parameter Estimation						
Masters	89	29.31[1] (73.28%)	4.77	109	38.11[4] (63.52%)	8.03
Nonmasters	911	14.18[2] (35.45%)	4.32	891	16.31[5] (27.18%)	5.70
Combined	1000	15.53[3] (38.83%)	6.13	1000	18.69[6] (31.15%)	9.06
Validation						
Masters	62	29.15[1] (72.88%)	4.25	64	39.00[4] (65.00%)	8.02
Nonmasters	610	14.15[2] (35.38%)	4.49	608	16.22[5] (27.03%)	6.03
Combined	672	15.53[3] (38.83%)	6.23	672	18.39[6] (30.65%)	9.15

[1, 2, 3, 4, 5, 6] Pairwise contrasts of mean test scores in corresponding groups and subgroups.

contrasts 1, 2, and 3 (see Table 1) differed significantly. Means of reading scores were compared similarly. Contrasts 4, 5, and 6 in Table 1 were not found to be statistically different.

The parameter estimation group and the validation group appeared to be essentially the same with respect to the two subtests.

Tables 2 and 3 show, respectively, the results of the listening test re-enactments and of the reading test re-enactments for the 672 examinees in the validation group.

Decision accuracy is the percentage of decisions made by each adaptive method that were the same as those reached on the basis of actual total scores. In other words, it is the percentage of "hits". Decision accuracy is a measure of how effective an adaptive testing method is, when compared with the administration of the whole test. Average length is the average number of test items required to reach a mastery or nonmastery decision under each adaptive testing method. It is a measure of the efficiency of an adaptive testing method, when compared with total test administration.

Decision Accuracy

The accuracy, or effectiveness, of an adaptive method was compared with the theoretical accuracy expectation. Since α and β were each set to 0.025 for EXSPRT, and 0.95 confidence interval of the examinee's θ was used with AMT to reach mastery or nonmastery conclusion, the overall *a priori* theoretical error rates (for false mastery decisions and false nonmastery decisions

Table 2
Efficiency and Accuracy of AMT, EXSPRT-I, and EXSPRT-R (40-item Listening Test)

Adaptive Method	Masters (n=62)		Nonmasters (n=610)		Total (n=672)	
	Mean Length (S.D.)	Accuracy	Mean Length (S.D.)	Accuracy	Mean Length (S.D.)	Accuracy
AMT	21.89 (16.96)	72.58%*	9.69 (10.92)	99.18%	10.81 (12.12)	96.73%
EXSPRT-I	11.68 (7.08)	91.94%	8.17 (7.19)	91.48%*	8.49 (7.25)	91.52%*
EXSPRT-R	11.95 (8.82)	93.55%	10.98 (8.46)	95.57%*	11.07 (8.50)	95.39%

*Percent accuracies that differ significantly from the *a priori* error rates ($\alpha=\beta=0.025$ for EXSPRT, and a 0.95 confidence interval was used with AMT).

Table 3
Efficiency and Accuracy of AMT, EXSPRT-I, and EXSPRT-R (60-item Reading Test)

Adaptive Method	Masters (n=64)		Nonmasters (n=608)		Total (n=672)	
	Mean Length (S.D.)	Accuracy	Mean Length (S.D.)	Accuracy	Mean Length (S.D.)	Accuracy
AMT	28.92 (26.57)	82.81%*	9.26 (14.61)	99.01%	11.13 (17.11)	97.47%
EXSPRT-I	8.48 (3.90)	95.31%	6.17 (8.95)	91.78%*	6.39 (8.62)	92.11%*
EXSPRT-R	11.45 (8.72)	96.88%	12.55 (10.46)	94.08%*	12.44 (10.31)	94.35%

*Percent accuracies that differ significantly from the *a priori* error rates ($\alpha=\beta=0.025$ for EXSPRT, and a 0.95 confidence interval was used with AMT).

combined) was 5%. The *a priori* error rate for false mastery decisions was 2.5%, and that for false nonmastery decisions was also 2.5%.

For each observed accuracy percentage reported in Tables 2 and 3, a 0.95 confidence interval was constructed. The confidence interval was then compared with the theoretical "hit" rate (one minus the theoretical error rate). Statistically significant deviation from the theoretical hit rate was concluded when the upper end of the confidence interval fell below the theoretical hit rate.

Tables 2 and 3 show the same pattern of statistically significant deviations of decision accuracies. In both subtests, AMT was the least accurate in making mastery decisions, while EXSPRT-R was the most accurate. But AMT was the most accurate in making nonmastery decisions, whereas EXSPRT-I was the least accurate. Furthermore, in both subtests, AMT made significantly more false mastery decisions than theoretically expected. EXSPRT showed a reverse pattern—both EXSPRT-I and EXSPRT-R made significantly more false nonmastery decisions. Overall, AMT was the most accurate in making mastery and nonmastery decisions, whereas EXSPRT-I was the least accurate. Total AMT and EXSPRT-R decision accuracies were statistically within the expected 0.95 level in both subtests.

It should be noted that even though EXSPRT-I and EXSPRT-R nonmastery decision accuracies deviated significantly from the theoretically expected 0.975 level in both subtests, they were

nonetheless higher than the observed AMT mastery decision accuracies. Decision accuracies higher than 90% are oftentimes acceptable in many testing situations. In fact, all significantly deviated EXSPRT-I and EXSPRT-R decision accuracies in Tables 2 and 3 were within the 90% level.

Frick, Luk and Tyan (1987) demonstrated that AMT and SPRT decision accuracies tended to agree better with theoretical expectations when the distribution of examinees' scores was flatter. Both distributions of listening scores for the validation group and the parameter estimation group were unimodal, leptokurtic, and negatively skewed, with modes at 15 (37.5%) and 13 (32.5%) respectively. The distributions of reading scores for the two groups were also unimodal, leptokurtic, and negatively skewed. The mode of reading scores was 17 (28.33%) for the validation group, and 16 (26.67%) for the parameter estimation group. The lepto-kurtosis of the distributions may partly account for the observation that some EXSPRT-I and EXSPRT-R decision accuracies were not statistically significant, even though they appeared to be relatively high.

Efficiency

Tables 2 and 3 show that in both subtests, EXSPRT-I required the least number of test items to reach a mastery or nonmastery decision for nonmasters, followed by AMT, and then by EXSPRT-R. The same pattern was observed in the total groups in both subtests. However, a different pattern of test lengths was observed in the two master groups—AMT required more items than

EXSPRT-R, while EXSPRT-I was still the method that required the least number of items.

Differences between the mean test lengths of the three adaptive methods were tested for each of the three grouping conditions (masters, nonmasters, and total) in each of the two subtests. Mean test lengths were tested by conducting repeated measures ANOVAs. Statistically significant differences were observed in all six cases. However, the sphericity assumption was violated in all cases. A *post-hoc* comparison procedure for this kind of situation was suggested by Marascuilo and Levin (1983, pp.373-381) and used in the present study.

Of the nine possible pairwise contrasts of listening test lengths, only two were found to be NOT significant: the difference between EXSPRT-I test length (11.68) and EXSPRT-R test length (11.95) for masters, and that between AMT (10.81) and EXSPRT-R (11.07) for the total group. Only one pairwise contrast of reading test lengths was found to be NOT significant: the difference between AMT (11.13) and EXSPRT-R (12.44) for the total group.

It should be noted that AMT and EXSPRT-I appeared to require less test items to reach a decision for nonmasters than for masters. T-tests revealed that AMT and EXSPRT-I did in fact require significantly less listening test items, and also significantly less reading test items, for nonmasters than for masters. No statistically significant differences in EXSPRT-R test lengths between masters and nonmasters were observed.

Frick et al. (1989) noted that the more test items there are available at different achievement levels, the more efficient and the more accurate AMT and EXSPRT-I can be. AMT and EXSPRT-I should work better with larger item pools. A small item pool may not contain as many informative and discriminating items as a large item pool. So it is more likely to run out of items that can provide much additional information when the pool is small. Results in the present study were consistent with this hypothesis. AMT and EXSPRT-I required a smaller **proportion** of items to reach decisions in the 60-item reading test than in the 40-item listening test.

Conclusion

Overall, it can be concluded that EXSPRT-I appears to be a strong, viable alternative to the three-parameter AMT model as an adaptive mastery testing method. EXSPRT is less complicated conceptually and mathematically, and is more practical in the sense that it does not require a large parameter estimation sample, as was demonstrated in Frick et al.'s (1989) studies.

Although EXSPRT-I decision accuracies were not as high as theoretically expected, they were reasonably high for many practical purposes. Had the distributions of examinees' test scores been less leptokurtic and less skewed than those observed in the present study, decision accuracies could have been in closer agreement with the theoretically expected levels. More research is needed to further investigate this hypothesis.

Results from the present study suggest that EXSPRT-I can be more efficient than the three-parameter AMT. Specifically, EXSPRT-I was more accurate and more efficient in making mastery decisions than AMT. However, it was less accurate but more efficient than AMT for nonmastery decisions.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord and M. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley

Brown, J., and Weiss, D. (1977). *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Ferguson, R. L. (1969). *Computer-assisted criterion-referenced measurement* (Working Paper No. 41). Pittsburg: University of Pittsburg, Learning Research and Development Center.

Frick, T. (1990). A comparison of three decision models for adapting the length of computer-based mastery tests. *Journal of Educational Computing Research, 6*, 479-513.

Frick, T., Luk, H., and Tyan, N. (1987). *A comparison of three adaptive decision-making methodologies used in computer-based instruction and testing* (Final Report, Proffitt Foundation). Bloomington, IN: Indiana University, School of Education.

Frick, T., Plew, G., and Luk, H. (1989, March). *EXSPRT: An expert systems approach to computer-based adaptive testing*. Paper presented at the Annual Conference of the American Educational Research Association, San Francisco, CA.

Hambleton, R. K., and Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing*. New York: Academic Press.

Kingsbury, G., and Weiss, D. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. Weiss (Ed.), *New horizons in testing*. New York: Academic Press.

Marascuilo, L., and Levin, J. (1983). *Multivariate statistics in the social sciences: A researcher's guide*. Monterey, CA: Brooks/Cole.

Owen, R. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Weiss, D., and Kingsbury, G. (1984). Application of computerized adaptive testing to education problems. *Journal of Educational Measurement, 21,* 361-375.

Wingersky, M. S., Barton, M. A., and Lord, F. M. (1982). *Logist 5, Version 1.0* [Computer program]. Princeton, N. J.: Educational Testing Service.

APPENDIX 1

Birnbaum's (1968) item information formula :

$$I_i(\theta) = \frac{(1-c_i)\,d^2 a_i^2\,[\text{logistpdf}(a_i, b_i, d, \theta)]^2}{\text{logistpdf}(a_i, b_i, d, \theta) + c_i\,[\text{logistcdf}(a_i, b_i, -d, \theta)]^2}$$

where

$$\text{logistpdf}(a_i, b_i, d, \theta) = \text{logistic probability density function}$$

$$= \frac{\exp(da_i(\theta - b_i))}{[\,1 + \exp(da_i(\theta - b_i))\,]^2}$$

$$\text{logistcdf}(a_i, b_i, -d, \theta) = \text{logistic cumulative density function}$$

$$= \frac{\exp(\,(-d)a_i(\theta - b_i)\,)}{1 + \exp(\,(-d)a_i(\theta - b_i)\,)}$$

Owen's Bayesian scoring algorithm :

(A) $E(\theta|1)$ = estimated θ given a correct response

$$= M_0 + \frac{\left[\,\frac{(1-c_i)V_0}{((1/a_i^2)+V_0)^{0.5}}\,\right]\phi(D)}{A}$$

where M_0 = prior θ estimate

V_0 = prior variance estimate of θ

$$\phi(D) = \left[\,\frac{1}{(2\pi)^{0.5}}\,\right]\left[\,\exp\left(\,-(D^2)/2\,\right)\,\right]$$

$$D = \frac{(\,b_i - M_0)}{((1/a_i^2)+V_0)^{0.5}}$$

$$A = c_i + (1-c_i)\,\Phi(-1.7D)$$

$$\Phi(x) = e^x \,/\, (1 + e^x)$$

(B) $E(\theta|0)$ = estimated θ given an incorrect response

$$= M_0 - \frac{\left[\dfrac{V_0}{((1/a_i^2)+V_0)^{0.5}}\right]\phi(D)}{\Phi(1.7D)}$$

(C) $V(\theta|1)$ = estimated θ variance given a correct response

$$= V_0\left\{1 - \left[\frac{(1-c_i)}{1+(1/(a_i^2V_0))}\right]\phi(D)\left[\frac{\left[\dfrac{(1-c_i)\phi(D)}{A}\right] - D}{A}\right]\right\}$$

(D) $V(\theta|0)$ = estimated θ variance given an incorrect response

$$= V_0\left\{1 - \left[\frac{\phi(D)}{1+(1/(a_i^2V_0))}\right]\left[\frac{\left[\dfrac{\phi(D)}{\Phi(1.7D)}\right] + D}{\Phi(1.7D)}\right]\right\}$$

(E) 0.95 confidence interval of θ is defined such that:

$$[E(\theta) - 1.96V(\theta)^{0.5}] \leq \theta \leq [E(\theta) + 1.96V(\theta)^{0.5}]$$

where $E(\theta)$ = θ estimate using Owen's algorithm

$V(\theta)$ = θ estimate using Owen's algorithm